UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Gustavo Vellasco Almeida de Lima

Análise de conjuntos de dados usados em métodos de detecção de anomalias em redes de computadores

Uberlândia, Brasil

2017

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Gustavo Vellasco Almeida de Lima

Análise de conjuntos de dados usados em métodos de detecção de anomalias em redes de computadores

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Rodrigo Sanches Miani

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2017

Gustavo Vellasco Almeida de Lima

Análise de conjuntos de dados usados em métodos de detecção de anomalias em redes de computadores

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 18 de dezembro de 2017.

Prof. Dr. Rodrigo Sanches Miani
Orientador

Prof. Dr. Rafael Pasquini

Prof. Luiz Cláudio Theodoro

Uberlândia, Brasil
2017

Dedico aos meus pais, minha namorada Mariana e aos meus amigos que, com muito carinho e apoio, não mediram esforços para que eu concluísse esta etapa da minha vida.

Agradecimentos

À Deus, por possibilitar a realização de mais um sonho, por me guiar em todas as minhas escolhas e sempre estar ao meu lado nos momentos de alegria e tristeza. Sem ele, não teria fé, não conseguiria acreditar que tudo isso seria possível.

Agradeço aos meus pais pela dedicação e esforço para minha criação, que tudo fizeram para que eu tivesse a melhor educação possível e ingressasse em uma instituição de ensino renomada, como é a Universidade Federal de Uberlândia.

À minha namorada pelo apoio, compreensão e companheirismo, que durante todos esses anos e principalmente durante o final da minha graduação, foi extremamente importante.

Ao meu orientador Prof. Dr. Rodrigo Sanches Miani, por ter aceito ser meu orientador na realização do trabalho de conclusão de curso e também pelas revisões e discussões que influenciaram diretamente neste trabalho, sendo fundamental para realização desta conquista.

Aos amigos que fiz durante a faculdade e que levarei para a vida.

"A persistência é o caminho do êxito." Charles Chaplin

Resumo

Com a constante evolução das soluções tecnológicas, a Internet se tornou um dos principais meios de comunicação, possuindo cada vez mais importância no âmbito comercial e no cotidiano das pessoas. A complexidade e a utilização em larga escala das redes de computadores dificultam o monitoramento e o controle, tornando comum a ocorrência de anomalias. Existem diversos trabalhos/ferramentas que analisam e detectam tais anomalias, porém há um problema associado à forma como os métodos propostos são validados. O objetivo deste trabalho é verificar de que modo tais métodos encontrados na literatura têm sido validados. Os resultados encontrados mostram que a maioria dos trabalhos analisados ainda usam conjuntos de dados sintéticos e obsoletos para a validação dos métodos de detecção de anomalias, porém, trabalhos mais recentes tendem a validar as técnicas propostas usando conjuntos de dados reais.

Palavras-chave: Detecção de Anomalias, Gerência de Redes, KDD, DARPA, Metodologia, Conjunto de Dados.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BGP	*Border Gateway Protocol*
DARPA	*Defense Advanced Research Projects Agency*
DBSCAN	*Density Based Spatial Clustering of Application*
DDoS	*Distributed Denial of Service*
DoS	*Denial of Service*
EWMA	*Exponentially Weighted Moving Average*
F-CBCT	*Fuzzified Cuckoo based Clustering Technique*
FTP	*File Transfer Protocol*
GLR	*Generalized Likelihood Ratio*
HTTP	*HyperText Transfer Protocol*
ICMP	*Internet Control Message Protocol*
ISCX	*Information Security Center of Excellence*
IDS	*Intrusion detection System*
IMAP	*Internet Message Access Protocol*
IMAPIT	*Integrated Measurement Analysis Platform for Internet Traffic*
INBOUNDS	*Integrated Network Based Ohio University Network Detective Service*
IP	*Internet Protocol*
IPv4	*Internet Protocol version 4*
IPv6	*Internet Protocol version 6*
ISP	*Internet Service Provider*
KDD	*Knowledge Discovery and Data Mining*
KNN	*K-nearest-neighbor*
LAMS	*Local Adaptive Multivariate Smoothing*

MIB	*Management Information Base*
NIDS	*Network Intrusion Detection Systems*
N/I	*Não Informado*
PCA	*Principal Component Analysis*
PDU	*Protocol Data Unit*
PKI	*Public Key Infrastructure*
PMC	*Program Management Console*
POP3	*Post Office Protocol version 3*
PTF	*Passive TCP/IP Fingerprinting*
RAD	*Registry Anomaly Detection*
RNN	*Recurrent Neural Network*
SNMP	*Simple Network Management Protocol*
SMTP	*Simple Mail Transfer Protocol*
SSH	*Secure Shell*
SOFM	*Self-Organized Feature Map*
SVM	*Support Vector Machines*
TCP	*Transmission Control Protocol*
UDP	*User Datagram Protocol*
UNSW	*University of New South Wales*
VPN	*Virtual Private Network*

Sumário

1 Introdução

Com a constante evolução das soluções tecnológicas e do mundo, a Internet se tornou um dos principais meios de comunicação, possuindo cada vez mais importância no âmbito comercial e no cotidiano das pessoas. Tal fato contribui para o crescimento exponencial do acesso à Internet, que exige cada vez mais investimentos em monitoramento e controle das redes de computadores, além de ampliações para suportar todo o crescente tráfego de dados.

A complexidade e a utilização em larga escala das redes de computadores dificultam o monitoramento e o controle, tornando comum a ocorrência de anomalias (ZARPELÃO, 2010). Anomalias podem ser definidas como desvios em relação a um comportamento padrão em redes de computadores, segundo Perlin, Nunes e Kozakevicius (2011), que podem ocasionar degradação na qualidade dos serviços ou até mesmo interrupção total deles, sendo principalmente causados por atividades maliciosas ou defeitos.

Nesse contexto, entende-se por atividades maliciosas, ataques que normalmente são arquitetados por agentes que visam romper as barreiras de segurança da rede, comprometendo seu desempenho (THOTTAN; JI, 2003). No caso das anomalias causadas por defeitos, elas podem surgir de falhas em elementos de rede ou até mesmo por vandalismos em cabos ópticos, onde trafegam os dados.

No último ano foram noticiadas na mídia algumas ocorrências que ilustram os impactos da interrupção da Internet para os usuários, como o caso em que um incêndio em um prédio da Algar Telecom ocorrido em maio de 2016, deixou parte da cidade de Uberlândia sem os serviços de telefonia fixa e Internet (G1, 2016a).

No que diz respeito às anomalias causadas por atividades maliciosas, no Brasil destaca-se uma matéria onde o site da Agência Nacional de Telecomunicações ficou indisponível após ser alvo de um ataque de negação de serviço distribuído (DDoS) (G1, 2016b). Há também um caso mais recente, ocorrido em maio de 2017, que afetou a operadora Telefônica da Espanha, que foi infectada pelo vírus *ransomware*, cujo objetivo é promover o sequestro de dados e cobrar um resgate para liberação das informações (CincoDías, 2017).

Garantir confidencialidade, integridade e disponibilidade aos sistemas nas redes de computadores é um desafio constante (CRISTINA; C, 2011). De modo a proteger os dados, serviços e recursos computacionais das redes, diversas ferramentas e trabalhos científicos foram desenvolvidos nas décadas de 90 e 2000. No que se refere às principais tecnologias de segurança, destacam-se aquelas voltadas para a segurança defensiva, como: antivírus, *firewall*, IDS *(Intrusion detection system)*, PKI *(Public key infrastructure)* e VPN *(Virtual private network)*, que visam proteger os dados, reduzindo a vulnerabilidade

e minimizando os riscos.

Com relação aos trabalhos voltados especificamente para a detecção de anomalias, destaca-se a caracterização do tráfego não malicioso da Internet nas pesquisas de Claffy (1994), seguido pelos trabalhos de Barford et al. (2002), He, Yu e Li (2008), Wang et al. (2008) e Zhani, Elbiaze e Kamoun (2008) que propõem metodologias e modelos de solução para o problema de identificação de anomalias. Por fim, encontram-se os trabalhos de Celenk et al. (2010), Zarpelão (2010), Gogoi et al. (2011) e Bartos, Rehak e Krmicek (2011) que abordam sobre a pesquisa, identificação, monitoramento e previsão de anomalias em redes de computadores.

Um problema relevante na área de detecção de anomalias está relacionado à validação dos métodos propostos nos trabalhos. Tais métodos costumam ser validados com o auxílio de conjuntos de dados onde geralmente são divididos em: conjuntos reais e conjuntos sintéticos. De acordo com Nehinbe (2011), é importante que tais métodos sejam validados da maneira mais realística possível, ou seja, usando conjuntos reais ou conjuntos que se assemelham aos conjuntos de dados reais.

O objetivo do trabalho é verificar de que forma tais métodos apresentados na literatura são validados, se usam conjuntos de dados reais e/ou sintéticos, além de analisar também os métodos de detecção de anomalias propostos em cada deles. Com esse intuito, 36 trabalhos foram analisados e classificados de acordo com o conjunto de dado de validação. Tais trabalhos foram extraídos do *survey "A survey of network anomaly detection techniques"* de Ahmed, Mahmood e Hu (2016).

Este trabalho é organizado da seguinte forma: o Capítulo 2 descreve os conceitos básicos sobre gerência de rede e detecção de anomalias nas redes de computadores, além de introduzir algumas técnicas de detecção de anomalias e trabalhos relacionados à área. O Capítulo 3 apresenta a metodologia adotada para encontrar os trabalhos a serem analisados seguindo uma sequência de passos. O Capítulo 4 traz os métodos de detecção de anomalias encontrados nos artigos e aborda sobre a definição dos conjuntos de dados de validação, apresentando as tabelas oriundas do estudo dos artigos. Por fim, o Capítulo 5 apresenta os números obtidos durante o desenvolvimento do trabalho e ressalta a importância das validações da maneira mais realística possível.

2 Revisão Bibliográfica

Este capítulo aborda os conceitos básicos relacionados à gerência de rede e detecção de anomalias nas redes de computadores. São apresentados no decorrer do capítulo algumas formas de monitoramento, as causas e tipos de anomalias, bem como as fontes de coleta de informações sobre a rede e os diferentes métodos existentes. Por fim, é apresentado um levantamento de soluções para detecção de anomalias baseada em trabalhos recentes.

2.1 Gerência de Rede

A gerência da rede de computadores é o meio pelo qual o administrador da rede, utilizando uma gerência centralizada (normalmente composta por uma solução de hardware e software), consegue monitorar todo o ambiente da rede, tornando o monitoramento mais efetivo (BUENO, 2012). Com esse recurso, é possível gerar dados históricos e estatísticos de cada evento, bem como atuar de maneira mais assertiva em momentos de falha ou indisponibilidade, reduzindo o tempo de recuperação.

Segundo Stallings (2005), um sistema de gerenciamento de redes é um conjunto de ferramentas para monitoramento e controle de rede, que é integrado nos seguintes sentidos:

- Deve conter uma interface única para o operador, que seja de fácil utilização e atenda a maioria das necessidades do gerenciamento cotidiano;

- Que a maior parte da implementação seja executada no dispositivo gerenciado.

O gerenciamento de redes de computadores, indiferente de desenho adotado pelo administrador ou desenvolvedor de software, trabalha basicamente sobre três pontos:

- **Coleta de dados (*pooling*):** este item atua na coleta de dados dos recursos gerenciados. A coleta é executada por um componente de hardware e software, que conforme um tempo determinado pelo administrador, executará uma série de coletas.

- **Análise:** este item executará a análise dos dados coletados e fará a inferência dos mesmos em relação aos parâmetros determinados pelo administrador, ou seja, se um dado valor recebido do processo de coleta de dados está dentro ou fora da normalidade esperada pelo administrador.

- **Ação:** este item atuará após a análise, neste ponto alguma ação pode ser executada. As ações podem ser, um alarme visual em uma interface de navegador de Internet, envio de e-mail ou o que mais for conveniente e suportado pela plataforma de gerenciamento.

  Os três pontos anteriormente descritos podem ser interpretados como um ciclo, conforme Figura 1.

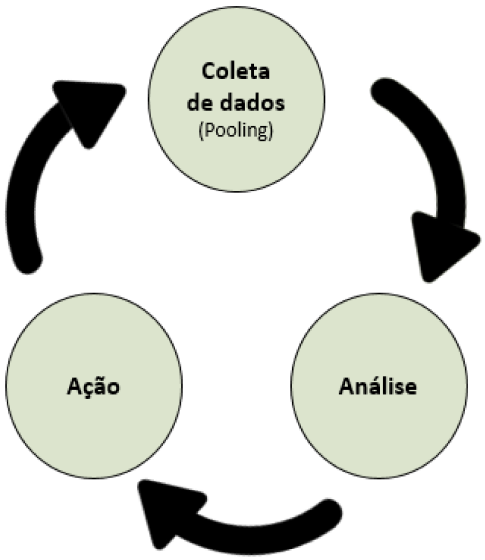


Figura 1 – Ciclo do Gerenciamento de Rede - Adaptado de: (BUENO, 2012)

2.1.1 Monitoramento via SNMP

Devido ao tamanho e complexidade das redes de computadores, surgiu a necessidade de ferramentas e protocolos para auxiliar na tarefa de gerência, tanto dos elementos quanto dos serviços. Basicamente, as ferramentas são compostas de gerentes e agentes e a comunicação entre eles é feita através de notificações de eventos, como ilustrado na Figura 2. Estes eventos são produzidos pelos elementos gerenciados, podendo indicar a ocorrência de uma anomalia (SAITO; MADEIRA, 2001).

Para poder gerenciar os recursos, os fabricantes de equipamentos de redes adotaram vários conjuntos de padrões para a operação de programas gerenciadores. Um conjunto de padrões popular, o SNMP *(Simple Network Management Protocols)*, serve como exemplo para todos os sistemas. Sob a arquitetura SNMP, pequenos programas de gerenciamento, conhecidos como agentes, são executados em um processador especial contido em uma variedade de dispositivos ligados à rede. Estes programas monitoram os dispositivos e coletam dados estatísticos em um formato conhecido como MIB *(Management Information Base)*. Um programa central, conhecido como PMC *(Program Management Console)* ordena os agentes em uma base regular e descarrega o conteúdo dos seus MIBs.

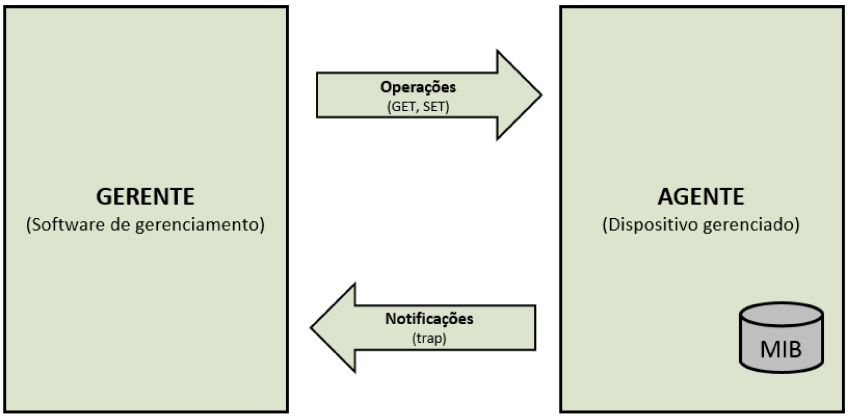


Figura 2 – Operações Básicas do SNMP - Adaptado de: (BUENO, 2012)

A comunicação entre agentes e gerentes em uma rede SNMP é baseada nas especificações das chamadas PDU *(Protocol Data Unit)*. A primeira versão do SNMP - SNMPv1 - definiu cinco PDUs núcleo, conforme abaixo:

- ***get-request-PDU:*** mensagem enviada pelo gerente ao agente solicitando o valor de uma variável;
- ***get-next-request-PDU:*** mensagem utilizada pelo gerente para solicitar o valor da próxima variável depois de uma ou mais variáveis que foram especificadas;
- ***set-request-PDU:*** mensagem enviada pelo gerente ao agente para solicitar que seja alterado o valor de uma variável;
- ***get-response-PDU:*** mensagem enviada pelo agente ao gerente, informando o valor de uma variável que lhe foi solicitado;
- ***trap-PDU:*** mensagem enviada pelo agente ao gerente, informando um evento ocorrido.

Uma vez que quatro, das cinco mensagens SNMP são do tipo pergunta e resposta, o protocolo SNMP usa o protocolo de transporte UDP *(User Datagram Protocol)*, da camada de aplicação, para facilitar o intercâmbio de informação entre os dispositivos de rede.

Além de ter sido projetado para operar sob UDP (um protocolo não orientado a conexão), o próprio SNMP também é um protocolo não orientado a conexão, sendo cada troca de mensagens uma transação diferente entre o agente e a estação de gerenciamento.

Cada estação de gerenciamento, como também o agente, devem implementar os protocolos SNMP e, por consequência UDP e IP *(Internet Protocol)* para poderem se comunicar. Tal imposição exclui do processo de gerenciamento dispositivos que não suportam parte dos protocolos TCP/IP *(Transmission Control Protocol)*, ou que, apesar

de implementarem o TCP/IP para suportar suas aplicações, não desejam adicionar mais carga ao seu sistema com o suporte ao protocolo SNMP.

Para que tais dispositivos também possam ser gerenciados, criou-se o conceito de agente *proxy*, demonstrado na Figura 3. Desta forma, este agente, que sabe se comunicar usando SNMP, responderia em nome do dispositivo que ele representa, e passaria o resultado da comunicação para o dispositivo de acordo com o protocolo que ele entende. Para tanto, este agente possui uma função de mapeamento que recebe as informações do dispositivo que ele representa e transforma as informações coletadas em mensagens SNMP e vice-versa.

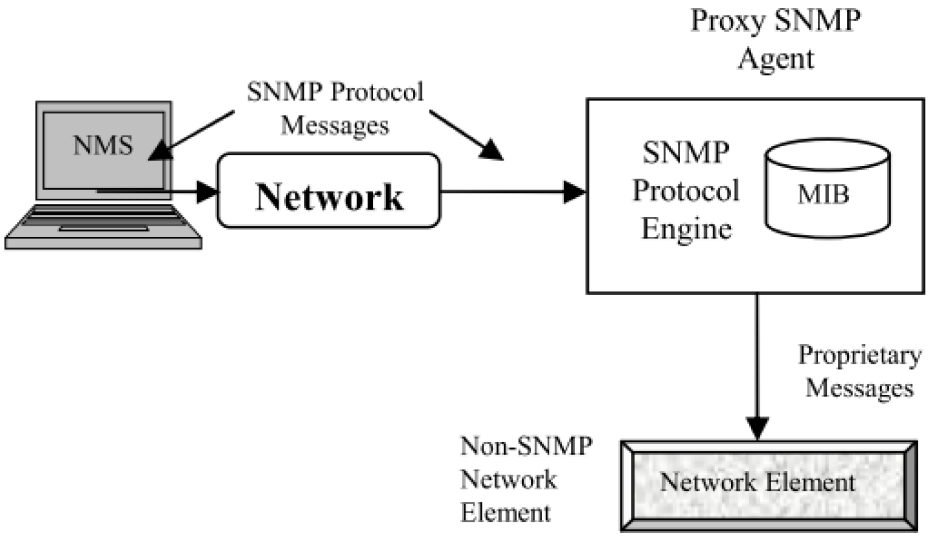


Figura 3 – SNMP Proxy Agent - Extraído de: (CHAVAN; MADANAGOPAL, 2008)

2.1.2 Monitoramento via ICMP

O programa *ping (Packet Internet Network Grouper)* é, talvez, a ferramenta de diagnóstico mais usada entre os profissionais da área. Originalmente criado para IPv4 *(Internet Protocol version 4)*, o ping foi estendido para se acomodar ao IPv6 *(Internet Protocol version 6)* (COMER, 2015).

Basicamente o *ping* envia pacotes ICMP *(Internet Control Message Protocol) echo request* para determinado *host*, este por sua vez, ao receber o pacote imediatamente retorna um pacote de *echo reply* para o *hosts* de origem, conforme ilutrado na Figura 4. É importante ressaltar que nem todos os *hosts* ou segmentos de rede permitem este tipo de teste (ABREU; PIRES, 2004).

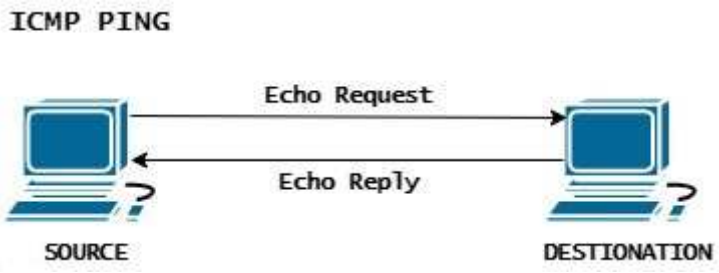


Figura 4 – Mensagens ICMP - Requisição e Resposta

Quando é obtido sucesso no teste de conectividade, como pode ser visualizado na Figura 5, opções mais avançadas podem ser usadas para mais testes, para que se aproximem das necessidades da aplicação que a rede deve suportar, como: definir o tamanho do pacote, desabilitar a opção de fragmentação, definir o tempo de *time-out* a ser observado ou quantidade de pacotes a ser enviado, entre outra opções.

```
C:\Users\gustavov>ping ufu.br

Disparando ufu.br [200.19.146.101] com 32 bytes de dados:
Resposta de 200.19.146.101: bytes=32 tempo=207ms TTL=61
Resposta de 200.19.146.101: bytes=32 tempo=450ms TTL=61
Resposta de 200.19.146.101: bytes=32 tempo=3ms TTL=61
Resposta de 200.19.146.101: bytes=32 tempo=7ms TTL=61

Estatísticas do Ping para 200.19.146.101:
    Pacotes: Enviados = 4, Recebidos = 4, Perdidos = 0 (0% de
             perda),
Aproximar um número redondo de vezes em milissegundos:
    Mínimo = 3ms, Máximo = 450ms, Média = 166ms
```

Figura 5 – Teste de ping

Caso seja observada perda parcial ou total de pacotes no teste com ping, é necessário determinar a causa de tal fato. Duas grandes causas de perda de pacote são: colisões ou descarte de pacotes por um dispositivo de rede por algum motivo desconhecido.

Em caso de falhas no teste com o *ping*, pode-se utilizar o *traceroute* para determinar até que ponto da rede o tráfego consegue fluir, ou seja, ter uma visão mais detalhada do ponto de falha na comunicação entre os dois *hosts* testados.

2.2 Detecção de Anomalias

Segundo Zarpelão (2010), existem diferentes formas de coletar informações de uma rede, e a escolha por uma delas passa por questões como a quantidade de recursos disponíveis para realizar o monitoramento e o tipo de problemas que se deseja identificar. A eficiência na detecção de anomalias está relacionada à escolha de um conjunto apropriado de dados para análise, que retrate de maneira fiel as características do funcionamento da rede monitorada. Os tipos de anomalias que serão detectadas dependem das caracterís-

ticas dos dados utilizados na detecção (ESTEVEZ-TAPIADOR; GARCIA-TEODORO; DIAZ-VERDEJO, 2004) (THOTTAN; JI, 2003).

No decorrer desta sessão serão apresentados alguns métodos de detecção de anomalias encontrados na literatura.

2.2.1 Detecção de anomalias usando estatísticas

Nos métodos estatísticos de detecção de anomalia, o sistema observa a atividade dos elementos e gera perfis para representar seu comportamento. Normalmente, dois perfis são mantidos para cada caso: o perfil atual e o perfil armazenado. À medida que os eventos de rede são processados, o sistema atualiza o perfil atual e calcula periodicamente uma pontuação de anomalia comparando-o com o perfil armazenado, usando uma função de anormalidade de todas as medidas dentro do perfil. Se a pontuação de anomalia for superior a um determinado limite, o sistema gera um alerta (ZHANG; YANG; GENG, 2009).

A detecção de anomalias através do método estatístico possui uma série de vantagens. Em primeiro lugar, esses sistemas não requerem conhecimento prévio de falhas ou ataques próprios. Além disso, as abordagens estatísticas podem fornecer uma notificação precisa de atividades que normalmente ocorrem em longos períodos de tempo. No entanto, os esquemas estatísticos de detecção de anomalia também apresentam inconvenientes, pois pode ser difícil determinar os limiares que equilibrem a probabilidade de falsos positivos com a probabilidade de falsos negativos. Além disso, os métodos estatísticos precisam de distribuições estatísticas precisas, mas nem todos os comportamentos podem ser modelados usando métodos puramente estatísticos (ZARPELÃO, 2010).

2.2.2 Detecção de anomalias baseado nas assinaturas das anomalias

Na detecção de anomalias baseada em assinaturas, as anomalias são modeladas de forma que suas principais características sejam levantadas e seja construída uma assinatura, que será armazenada em uma base de informações do sistema. Durante o monitoramento do tráfego de rede, o sistema busca comportamentos que tenham os mesmos atributos das assinaturas contidas na base de informações. Caso sejam encontradas situações semelhantes às descritas nas assinaturas, alarmes são enviados ao administrador de rede (ZARPELÃO, 2010).

A principal vantagem deste método é que ataques presentes na base de informações podem ser detectados de maneira eficaz, com baixas taxas de falsos positivos. A assinatura reúne uma série de eventos específicos para definir uma falha, permitindo que o problema seja facilmente diagnosticado. Outra importante vantagem reside no fato de que este tipo de sistema começa a proteger a rede logo após a sua implantação, pois não há período de

treinamento ou aprendizado sobre o funcionamento da rede.

Por outro lado, existe uma desvantagem onde os sistemas de detecção baseados em assinaturas não têm a capacidade de detectar falhas com características desconhecidas da sua base. Além disso, sistemas que se baseiam neste método consomem muitos recursos da equipe de administração de redes, já que a base que contém as assinaturas precisa ser constantemente atualizada com os novos ataques criados e novos elementos que foram inseridos no ambiente monitorado (PATCHA; PARK, 2007) (COLE, 2011).

2.2.3 Detecção de anomalias baseado na caracterização do comportamento normal

A detecção baseada na caracterização do comportamento normal, compara as informações coletadas da rede com as características das atividades consideradas normais, conforme a Figura 6. O perfil de normalidade é obtido após um estudo do comportamento prévio da rede. Uma situação é considerada anômala quando o seu grau de desvio em relação ao perfil de normalidade é significativo (COLE, 2011) (ESTEVEZ-TAPIADOR; GARCIA-TEODORO; DIAZ-VERDEJO, 2004) (PATCHA; PARK, 2007) (THOTTAN; JI, 2003) (PROENCA et al., 2006).

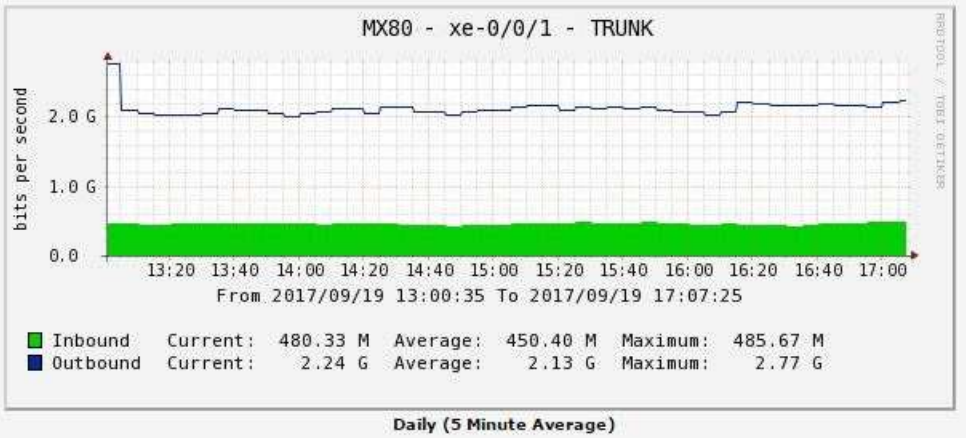


Figura 6 – Comportamento Normal

A principal vantagem deste método é a capacidade de detectar anomalias desconhecidas, ou seja, que não há histórico conhecido. Como a operação de detecção é baseada no conceito de normalidade e não na busca por comportamentos conhecidos e considerados anormais, um ataque com características desconhecidas, por exemplo, possivelmente será detectado, simplesmente por fazer com que o comportamento da rede se desvie de sua normalidade, conforme demonstrado na Figura 7.

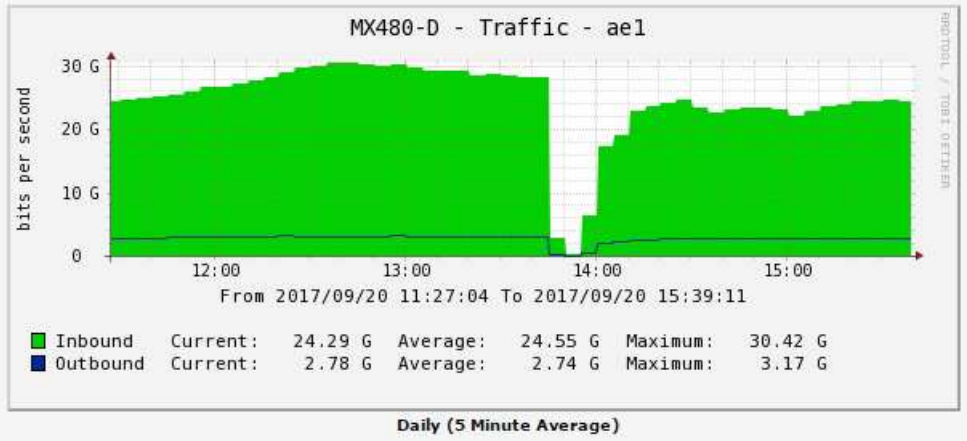


Figura 7 – Comportamento Anômalo

Pode-se considerar que a primeira desvantagem é a necessidade de um período de treinamento extenso para que os perfis de normalidade sejam aprendidos e construídos, além da alta porcentagem de falsos alarmes encontrados nestes sistemas. Muitas vezes as variações de comportamento são naturais nas redes, como a comutação de circuitos, e podem ser encaradas como desvios de comportamento por estes sistemas. A caracterização do tráfego é também uma questão de difícil solução, normalmente as estatísticas obtidas no monitoramento das redes apresentam características não estacionárias, que dificultam o estabelecimento de um padrão de operação. Uma modelagem de comportamento de baixa qualidade vai implicar em um sistema ineficiente.

2.3 Trabalhos Correlatos

Esta sessão tem por objetivo listar e descrever de maneira breve e sucinta os trabalhos relacionados à área de detecção de anomalias, juntamente com os conjuntos de dados usados para validação.

Trabalhos sobre detecção de anomalias em redes de computadores têm sido publicados por diferentes autores há quase três décadas, quando os primeiros estudos foram apresentados, como o de Anderson (1980). Diferentes conceitos e padrões da detecção de anomalias são encontrados nos trabalhos de Estevez-Tapiador, Garcia-Teodoro e Diaz-Verdejo (2004) e Lim e Jones (2008). Os dois trabalhos apresentam categorias para classificar os diferentes sistemas de detecção de anomalias. O trabalho apresentado por Estevez-Tapiador, Garcia-Teodoro e Diaz-Verdejo (2004) apresenta uma lista de sistemas relacionados à projetos de pesquisa, enquanto o trabalho de Lim e Jones (2008) expõe produtos comerciais.

Thottan e Ji (2003) propõem em seu trabalho um sistema de detecção de anomalias com o objetivo de realizar a correlação do comportamento de diferentes objetos

monitorados via SNMP para diminuir a taxa de falsos positivos. Os dados obtidos de cada objeto são coletados e organizados em séries temporais modeladas a partir da aplicação de um processo auto regressivo. O comportamento destas séries temporais é aplicado a um teste de hipótese baseado no método GLR *(Generalized Likelihood Ratio)* para detectar mudanças. Os resultados deste teste são organizados em vetores, que são correlacionados posteriormente com base nas características dos objetos SNMP. Os dados utilizados no trabalho foram obtidos de dois ambientes reais e operacionais, a rede corporativa da Lucent Technologies e a rede do *campus* Bell Laboratories.

O trabalho de Barford et al. (2002) apresenta técnicas de processamento de sinais em séries temporais, que são coletadas em MIBs e em registros de fluxos de dados. Para analisar as estatísticas do tráfego de rede obtidos através destas duas fontes, os autores desenvolveram o sistema IMAPIT *(Integrated Measurement Analysis Platform for Internet Traffic)* que aplica algoritmos de *wavelet* para decompor as séries temporais em conjuntos com diferentes frequências. O conjunto com as baixas frequências é formado por eventos de longa duração e os com médias e altas frequências trazem os eventos de curta duração. Desta forma, anomalias que são caracterizadas por eventos de longa duração, por exemplo, podem ser mais facilmente visualizadas no conjunto de baixas-frequências. Cada um destes conjuntos de frequências é analisado por um algoritmo denominado *deviation score*, que detecta as anomalias. Nesse artigo, foram utilizadas as informações de rede IP e dados SNMP coletados no roteador de borda da Universidade de Wisconsin, por um período de seis meses.

Lakhina, Crovella e Diot (2004) analisam estatísticas do tráfego de rede coletadas em registros de fluxos de pacotes do *backbone* Abilene, que liga as universidades e laboratórios de pesquisa da Internet em todo o continente americano. No trabalho é proposta uma técnica para diagnosticar anomalias, oferecendo não só a indicação de um evento, mas também a localização da origem desta anomalia. Toda a análise dos dados é baseada em uma técnica de mineração de dados conhecida como PCA *(Principal Component Analysis)*, capaz de separar as estatísticas coletadas em dois subgrupos: o subgrupo normal e o subgrupo anômalo. Um trabalho mais recente de Ringberg et al. (2007) demonstrou que é complexo aplicar esta técnica em ambientes de produção, já que o ajuste dos parâmetros levam à grandes variações na taxa de falsos positivos, prejudicando as análises e onerando os operadores durante a tratativa de uma falha. Os ambientes de produção utilizados neste trabalho foram as redes do *backbone* Abilene e do *backbone* Géant Network, que liga as redes nacionais de pesquisa e educação em toda a Europa.

No trabalho de Roughan et al. (2004) é apresentada uma proposta que explora a união de duas fontes de dados para melhorar o desempenho do sistema: o protocolo de gerência SNMP e o protocolo de roteamento BGP *(Border Gateway Protocol)*. A partir da correlação entre desvios de comportamento encontrados nestas duas fontes de dados,

foi alcançada uma redução no índice de falsos positivos. A análise dos dados coletados nos objetos SNMP foi realizada com o método *Holt-Winters* e a análise dos dados do protocolo BGP foi realizada com o método EWMA *(Exponentially Weighted Moving Average)*. Os dados analisados foram coletados de um ambiente real e operacional por mais de um ano, a rede IP do *backbone* do ISP *(Internet Service Provider)* tier-1.

Shon e Moon (2007) propõem em seu trabalho o uso de uma ferramenta comumente utilizada para reconhecimento de padrões e classificação, o SVM *(Support Vector Machines)*. Os dados usados para análise foram coletados dos ambientes de teste e produção do MIT Lincoln Labs. Primeiramente, os pacotes são coletados da rede e filtrados em tempo real pela ferramenta denominada PTF *(Passive TCP/IP Fingerprinting)*, que permite que pacotes mal formados sejam identificados e descartados. No conjunto de pacotes filtrados pelo PTF são aplicados dois processos: O primeiro visa determinar o perfil dos pacotes normais utilizando uma técnica de mineração de dados conhecida como SOFM *(Self-Organized Feature Map)*. Já o segundo processo utiliza algoritmos genéticos para selecionar quais campos dos pacotes apresentam maior probabilidade de evidenciar a ocorrência de anomalias. O resultado da execução destes processos é inserido na SVM para que seja efetuado o aprendizado e, nos próximos eventos, sejam detectadas anomalias.

No trabalho de Zarpelão (2010) é proposto um sistema de detecção de anomalias em redes de computadores baseado em três níveis de análise. O primeiro nível é responsável por comparar os dados coletados em um objeto SNMP com o perfil de operações normais da rede. O segundo nível correlaciona os alarmes gerados no primeiro nível de análise utilizando um grafo de dependências que representa as relações entre os objetos SNMP monitorados. O terceiro nível reúne os alarmes de segundo nível utilizando informações sobre a topologia de rede e gera um alarme de terceiro nível que reporta a propagação da anomalia pela rede. Os testes foram realizados na rede da Universidade Estadual de Londrina, utilizando situações reais. Os resultados apontaram que a proposta apresentou baixas taxas de falsos positivos combinadas a altas taxas de detecção. Além disso, o sistema foi capaz de correlacionar alarmes gerados para diferentes objetos SNMP em toda a rede, produzindo conjuntos menores de alarmes que ofereceram ao administrador de redes uma visão panorâmica do problema.

Papalexakis, Beutel e Steenkiste (2012) propõem aplicar um conjunto de técnicas emergentes de mineração de dados e de aprendizagem de máquinas ao problema de detecção de intrusão de rede, classificando as conexões como "normais"e "anômalas". Analisam a eficácia da utilização de dois algoritmos de *co-clustering* diferentes para dois *clusters* diferentes, marcando quais as medidas de conexão são indicadores fortes que para caracterizar um conjunto de dado anômalo. Os experimentos foram executados usando os algoritmos de *co-clustering* no conjunto de dados KDD *(Knowledge Discovery and Data Mining)* Cup 1999. Os autores acreditam que as idéias apresentadas neste trabalho podem

inspirar pesquisas para a detecção de anomalias em redes sociais, como a identificação de *spammers* e fraudadores.

É possível observar que nos trabalhos citados anteriormente, os conjuntos de validações utilizados variam entre reais e sintéticos. Cada tipo de conjunto possui suas vantagens e desvantagens. Por exemplo, conjuntos sintéticos são mais fáceis de serem obtidos e permitem a replicação de experimentos, enquanto os conjuntos reais podem fornecer situações mais desafiadoras e próximas daquelas em que o sistema de detecção de anomalia irá enfrentar. O objetivo do presente trabalho é fornecer uma análise sobre os tipos de conjuntos comumente usados para validação de métodos de detecção de anomalia.

3 Metodologia

A metodologia é compreendida como um nível aplicado para examinar, descrever e avaliar métodos e técnicas de pesquisa que possibilitam a coleta e o processamento de informações, visando o encaminhamento e a resolução de problemas e/ou questões de investigação (PROVDANOV; FREITAS, 2013).

Este trabalho consiste em uma pesquisa cientifica que tem como objetivo verificar de que modo tais métodos encontrados na literatura têm sido validados e quais tipos de conjuntos de dados estão sendo utilizados. Ao todo foram analisados 36 trabalhos, classificando-os de acordo com o conjunto de dado de validação. Tais trabalhos foram retirados do artigo *"A survey of network anomaly detection techniques"* de Ahmed, Mahmood e Hu (2016).

Existem diversos trabalhos publicados na literatura sobre o desenvolvimento de métodos de detecção de anomalias. O primeiro problema da pesquisa envolveu identificar um conjunto de trabalhos que serviriam como base para a análise. Para encontrar tais trabalhos foram seguidos os seguintes passos, conforme a Figura 8:

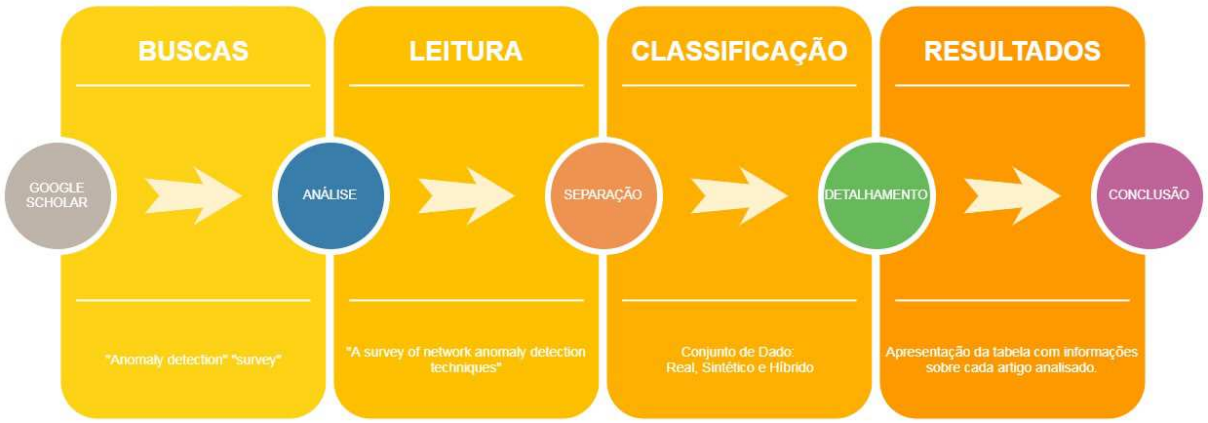


Figura 8 – Etapas do estudo

1. Realizar buscas no portal acadêmico Google Scholar;
2. Filtrar os trabalhos e ler cada um deles procurando por referências ao conjunto de dado usado para validação e técnica de detecção de anomalia utilizada;
3. Classificar os conjuntos de dados em reais e/ou sintéticos e também pela técnica de detecção de anomalia adotada;
4. Analisar os resultados.

O primeiro passo foi realizar a busca pela *string "Anomaly detection" "survey"* no Google Scholar, que retornou 36.300 artigos relacionado à consulta. O objetivo foi encontrar e separar os trabalhos relevantes que propõem métodos de detecção de anomalias. Foram pesquisados somente *surveys*, pois eles compreendem uma grande parte do estado da arte em uma determinada área, tornando a busca mais simples.

No segundo passo, foram avaliados os diversos *surveys* encontrados no passo anterior e destacados 23 deles. Com base no conteúdo de cada um, foi selecionado o trabalho "A survey of network anomaly detection techniques"de Ahmed, Mahmood e Hu (2016) por ser o mais novo dentre os encontrados. Neste *survey*, a classificação proposta pelos autores em relação ao método, engloba as técnicas mais comuns de detecção de anomalia e auxilia na organização dos trabalhos que serão analisados.

Já o terceiro passo, consistiu em identificar cada um dos trabalhos que propõem métodos/ferramentas de detecção de anomalias no *survey* escolhido de Ahmed, Mahmood e Hu (2016). Para isso, foram lidas as classificações destes artigos no *survey*, analisando as citações e separando somente as que estavam diretamente relacionadas a um trabalho de detecção de anomalia, que foi validado usando algum conjunto de dado. Por exemplo, o trabalho *"Enhancing big data security with collaborative intrusion detection"* de Tan et al. (2014) foi citado no *survey* dentro da classificação de *Information theory*, contudo, ao analisá-lo constatou-se que estava fora do escopo, por não estar relacionado à propostas/técnicas de detecção de anomalias que foram validadas com algum conjunto de dados, seja ele real ou sintético.

Nessa fase foram identificados 34 trabalhos publicados até 2014, onde 26 trabalhos estavam dentro do escopo da pesquisa e 8 trabalhos fora do escopo.

Com o intuito de recuperar trabalhos mais novos do que aqueles encontrados anteriormente, foi feita uma nova busca no Google Scholar com a *string* "network" "anomaly detection", refinando por artigos publicados a partir de 2015 e desconsiderando patentes. Essa nova busca gerou 10 novos artigos para análise, que serão apresentados no Capítulo 4 juntamente com os demais.

Por fim, a última etapa consiste em analisar os conjuntos de dados usados para a validação das propostas em cada um dos trabalhos selecionados. O detalhamento da análise, assim como a descrição dos resultados será realizada no Capítulo 4.

4 Resultados

O objetivo deste capítulo é apresentar os resultados dos passos elencados no Capítulo 3, juntamente com mais informações sobre os dois principais atributos usados para classificar os artigos: método de detecção e tipo de conjunto de dado.

As Tabelas 2, 3, 4, 5, 6, 7 e 8 apresentam os trabalhos selecionados com o auxílio da metodologia definida no Capítulo 3, ordenados por ano, contendo informações sobre o método de detecção utilizado, ano de publicação, tipo de conjunto de dado, e outras informações coletadas durante o estudo dos artigos.

4.1 Métodos de Detecção

De acordo com Ahmed, Mahmood e Hu (2016), existem quatro possíveis métodos de detecção de anomalias em redes de computadores: *classification-based*, *clustering-based*, *information theory* e *statistical*. Com a busca por artigos mais recentes, foi encontrado outro método, o *learning-based*. A seguir, cada um deles será brevemente explicado.

4.1.1 *Classification-based*

As técnicas baseadas em classificação dependem do extenso conhecimento dos especialistas de rede sobre as características dos ataques. Quando um especialista fornece detalhes das características do sistema de detecção, um ataque com um padrão conhecido pode ser detectado assim que for iniciado. Isso depende apenas da assinatura do ataque no sistema, tornando-o capaz de detectá-lo somente se a assinatura for fornecida anteriormente por um especialista de rede, deixando o sistema vulnerável a novos ataques que aparecem constantemente em diferentes versões. Mesmo que a assinatura de um novo ataque seja criada e incorporada ao sistema, a perda inicial pode ser insubstituível e o procedimento de reparo extremamente caro (AHMED; MAHMOOD; HU, 2016).

4.1.2 *Clustering-based*

A metodologia baseada em *clustering* refere-se a algoritmos de aprendizagem não supervisionados que não requerem dados pré-rotulados para extrair regras para agrupar instâncias de dados semelhantes (JAIN; MURTY; FLYNN, 1999). Embora existam diferentes tipos de técnicas de agrupamento, nos artigos estudados no Capítulo 3, as técnicas encontradas foram: *clustering regular* e *co-clustering*. De acordo com Ahmed, Mahmood e Hu (2016), a diferença entre eles é o modo de processamento de linhas e colunas. Técnicas regulares de agrupamento como *k-means* agrupam os dados considerando as linhas do

conjunto de dados, enquanto o *co-clustering* considera as linhas e colunas do conjunto de dados simultaneamente para produzir *clusters*.

4.1.3 *Information theory*

As técnicas de teoria da informação analisam o conteúdo da informação de um conjunto de dados usando diferentes medidas teóricas de informação, como: *Kolomogorov Complexity, entropy, relative entropy* e assim por diante. Tais técnicas são baseadas no seguinte pressuposto-chave: suposição (LEE, 2001).

4.1.4 *Learning-based*

De acordo com Patcha e Park (2007), a aprendizagem automática pode ser definida como a capacidade de um programa e/ou um sistema de aprender e melhorar seu desempenho em uma determinada tarefa ou grupo de tarefas ao longo do tempo. A aprendizagem de máquinas visa responder a muitas das mesmas questões estatísticas e de mineração de dados. No entanto, ao contrário das abordagens estatísticas que tendem a se concentrar na compreensão do processo que gerou os dados, as técnicas de aprendizado de máquinas se concentram na construção de um sistema que melhore seu desempenho com base em resultados anteriores. Por outras palavras, os sistemas baseados no paradigma de aprendizado de máquina têm a capacidade de mudar sua estratégia de execução com base em informações recém-adquiridas.

4.1.5 *Statistical*

Segundo Benso (2003), métodos estatísticos são empregados nos sistemas de detecção de intrusões, principalmente baseados na detecção de anomalias, para determinar a ocorrência de intrusões ou para o pré-processamento dos dados que serão utilizados por outros métodos, como redes neurais.

A vantagem deste método, de acordo com Ye et al. (2002), é a sua capacidade de tratar e representar explicitamente as variações e ruídos envolvidos nas atividades dos sistemas computacionais. Além disso, eles apresentam um bom desempenho computacional, bons índices de reconhecimento e boa escalabilidade, minimizando o tempo de resposta dos sistemas de detecção, aumentando a confiança no resultado do sistema e reduzindo gargalos de processamento.

4.2 Validações

Em seu trabalho, Balci (1997) define que validação é o ato de verificar se o modelo, dentro de seu domínio de aplicação, se comporta de maneira satisfatória de acordo com

os objetivos do estudo. Em outras palavras, a validação consiste em verificar se o modelo construído é o correto para a situação (CHRUN, 2011).

No contexto de trabalhos que tratam sobre o desenvolvimento de métodos para detecção de anomalias, é comum os pesquisadores elaborarem experimentos para validarem o método proposto. A elaboração do experimento passa, necessariamente, pela criação de um ambiente de rede com tráfego normal e tráfego anômalo. A seguir, serão apresentadas as três formas comumente usadas pelos pesquisadores para obter o tráfego de rede: uso de conjuntos de dados reais, conjuntos de dados sintéticos e conjuntos de dados híbridos.

4.2.1 Conjunto de dado real

Com a mudança contínua de comportamentos e padrões de rede, juntamente com a evolução das anomalias, torna-se necessário afastar-se de conjuntos de dados sintéticos e únicos, e usar conjuntos de dados gerados dinamicamente em ambientes reais, pois esses não refletem apenas as composições de tráfego e intrusões, mas também são modificáveis, extensíveis e reprodutíveis (SHARAFALDIN et al., 2017).

Exemplos de dados extraídos de ambientes reais são: *honeypots*, trafego de rede de empresas de telecomunicações, como exposto no trabalho de Deljac, Randi e Gordan (2015), ou até mesmo de uma universidade, como apresentado no trabalho de Fernandes et al. (2016), que utilizou os fluxos de IP de um *switch core* de rede da Universidade Estadual de Londrina, composto por cerca de 7000 dispositivos interconectados para validar a técnica proposta em seu trabalho.

De acordo com Shiravi et al. (2012), muitos desses conjuntos de dados são internos e não podem ser compartilhados devido a problemas de privacidade, outros são anônimos e não refletem tendências atuais ou eles não possuem certas características estatísticas.

É necessário mencionar também que com base na evolução do *malware* e mudanças contínuas nas estratégias de ataques e anomalias, os conjuntos de dados de referência precisam ser atualizados periodicamente, para que estejam o mais próximo possível da realidade.

4.2.2 Conjunto de dado sintético

Segundo Poojitha, Naveen e Jayarami (2010), existem duas formas de validar métodos de detecção de anomalias e intrusão: uma é criar a própria rede de simulação e coletar dados relevantes e a outra é usando conjuntos de dados existentes. A grande vantagem em usar conjuntos de dados existente é que os resultados podem ser comparados com outros na literatura.

Alguns dos conjuntos de dados sintéticos popularmente utilizados para validação de trabalhos referentes a detecção de anomalias são: conjunto de dados DARPA 1998, o

conjunto de dados DARPA 1999, o conjunto de dados KDD Cup 1999 e o conjunto de dados NSL KDD, que estão disponíveis no MIT Lincoln Laboratory.

Por exemplo, um dos conjuntos de dados muito utilizado é o DARPA, uma das primeiras coleções para análise de tráfego para detecção de intrusos. A criação desse conjunto foi uma iniciativa do grupo de tecnologias e sistemas cibernéticos do MIT Lincoln Laboratory, com o patrocínio da Agência de Projetos de Pesquisa Avançada de Defesa (DARPA) e da Força Aérea Americana (HAINES, 2017). O foco era mensurar a probabilidade de ataques e alarmes falsos em uma rede, onde as análises do tráfego da rede serviram como guia para muitos pesquisadores avaliarem as pesquisas desenvolvidas para a detecção de anomalias em redes de computadores.

4.2.3 Conjunto de dado híbrido

São classificados como conjuntos de dados híbridos, os trabalhos propostos na literatura que utilizam conjuntos de dados reais e sintéticos para validação dos métodos propostos no decorrer do artigo.

Um exemplo de trabalho que usou um conjunto de dados híbrido para validação das técnicas propostas é o de Lee (2001), onde é utilizado um conjunto de dado sintético popularmente conhecido, o DARPA 1999 e também um conjunto de dado real, que são os dados do sistema de *sendmail* da Universidade do Novo México.

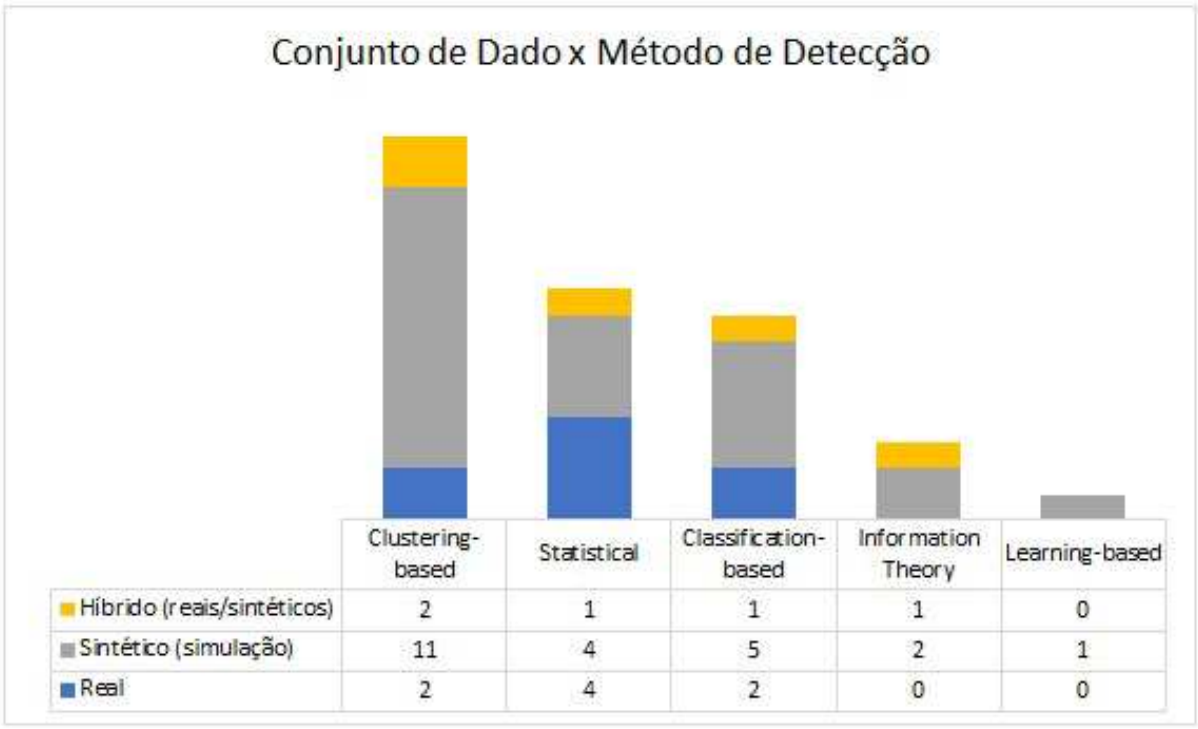
4.3 Síntese dos resultados

A análise das tabelas mostram que em sua grande maioria os trabalhos analisados são validados utilizando conjuntos de dados sintéticos, conforme ilustrado na Figura 9. Isso mostra que boa parte dos métodos de detecção de anomalia investigados podem não exibir o mesmo comportamento ao serem reproduzidos em um ambiente real.

De acordo com Haines (2017), muitas pesquisas se baseiam em análises de dados obsoletos, com conjuntos de dados antigos, alguns com quase duas décadas, como o DARPA 1998 e o KDD Cup 1999.

Na literatura, é possível encontrar alguns conjuntos de dados sintéticos mais atuais que podem ser usados para validação de métodos de detecção de anomalias, como o ISCX *(Information Security Center of Excellence)* IDS *Dataset* de 2012, que possui uma abordagem moderna e mais próxima de um ambiente real, onde a geração desse conjunto de dados foca em tráfego de pacotes HTTP, SMTP, SSH, IMAP, POP3 e FTP, conforme definido no trabalho de Shiravi et al. (2012). Há também outro conjunto de dado sintético ainda mais atual, o UNSW-NB15 de 2015, que reune mais de 100 GB de capturas de tráfego de rede coletados em dois momentos distintos: o primeiro no dia 22 de Janeiro

de 2015, com duração de 16 horas, e o segundo em 17 de Fevereiro de 2015, por mais 15 horas, de acordo com o trabalho de Moustafa e Slay (2016).



	Clustering-based	Statistical	Classification-based	Information Theory	Learning-based
Híbrido (reais/sintéticos)	2	1	1	1	0
Sintético (simulação)	11	4	5	2	1
Real	2	4	2	0	0

Figura 9 – Análise do método de detecção de anomalia e do conjunto de dado utilizado.

Contudo, nota-se na Figura 10 que a partir de 2015 houve um aumento na quantidade de trabalhos validados usando conjuntos de dados reais. De um total de 8 artigos validados utilizando dados reais, 5 deles são após o ano de 2014. Isso indica uma possível tendência de que nos trabalhos mais recentes, as validações das técnicas propostas sejam através de conjunto de dados reais.

Tais conjuntos de dados reais utilizados para validação dos trabalhos, variam desde o uso de tráfego de rede corporativa como o cedido pela empresa Lucent Technologies (THOTTAN; JI, 2003) e pela operadora T-HT Telecom Deljac, Randi e Gordan (2015) até conjuntos de dados cedidos por universidades, como o trabalho proposto em (FERNANDES et al., 2016) que usa dados do tráfego de rede da Universidade Estadual de Londrina. De modo geral, o tempo de coleta dos conjuntos de dados reais presentes nos trabalhos analisados variam de 2 meses à 1 ano e não estão disponíveis publicamente, podendo ser um dos motivos pelos quais alguns pesquisadores ainda utilizam conjuntos sintéticos, pela dificuldade de encontrar publicamente conjuntos de dados reais confiáveis.

	1996	1999	2000	2001	2002	2003	2005	2006	2007	2010	2011	2012	2013	2014	2015	2016	2017
Real	0	0	0	0	0	2	0	0	0	0	0	0	1	0	2	2	1
Híbrido (reais/sintéticos)	1	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0
Sintético (simulação)	0	1	0	2	2	5	1	1	1	1	1	2	0	2	1	1	2
TOTAL	1	1	1	3	2	7	1	1	1	1	1	2	2	2	3	4	3

Figura 10 – Análise do ano e do conjunto de dado utilizado.

Para avaliar com mais detalhes quais conjuntos de dados/tipos de conjuntos foram utilizados para validação das técnicas de detecção de anomalias, os trabalhos foram separados por ano e categorizados em grupos para facilitar a análise, conforme ilustrado na Tabela 1.

Tabela 1 – Divisão em grupos

Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**
1995 à 2000	2001 à 2005	2006 à 2010	2011 à 2015	À partir de 2016

No gráfico exposto na Figura 11, observa-se que a partir dos trabalhos agrupados no grupo 3, há um aumento no número de validações utilizando conjuntos de dados reais e redução das validações usando conjuntos de dados sintéticos, como o DARPA e o KDD do MIT Lincoln Laboratory, os mais encontrados durante a análise dos trabalhos. Lembrando que existem trabalhos que foram validados usando mais de um conjunto sintético e outros que foram validados usando conjuntos de dados reais e sintéticos, classificados como híbridos.

Como é possível notar, isso reforça o argumento de que há uma propensão de que com o passar do tempo novos trabalhos sejam validados usando algum conjunto de dado real. As razões que envolvem essa situação devem ser investigadas com mais cuidado, porém essa pode ser uma indicação da obsolescência dos conjuntos sintéticos disponíveis para os pesquisadores da área de segurança da informação. Novos métodos deveriam ser validados em conjuntos sintéticos robustos e atuais e também em ambientes reais.

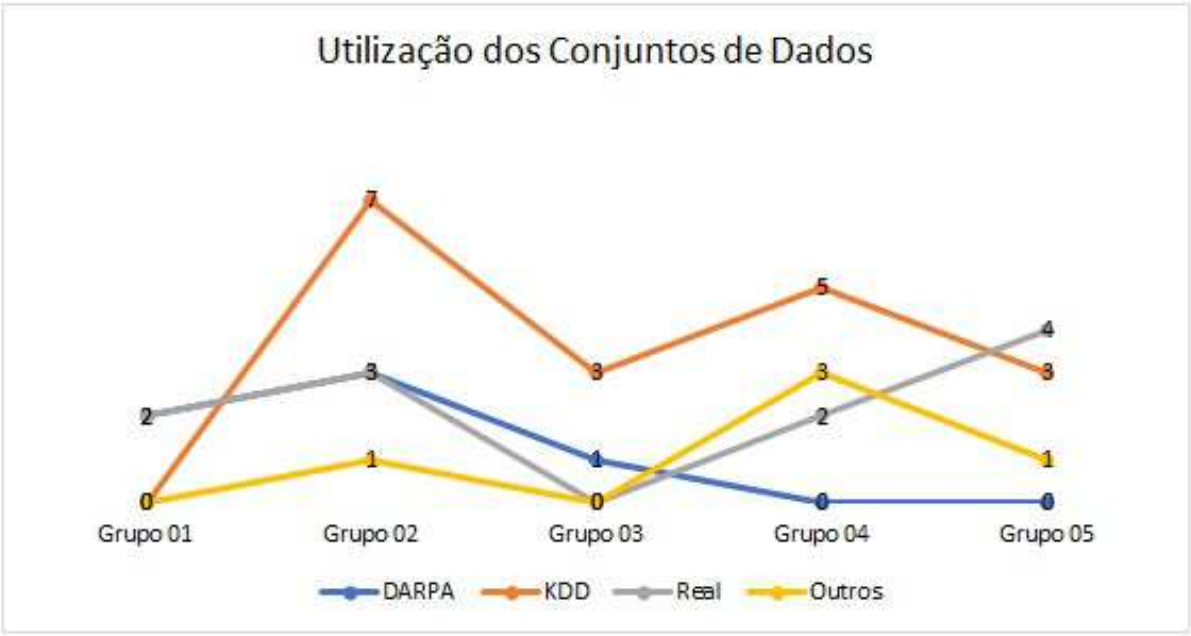


Figura 11 – Análise das validação usando alguns conjuntos de dados.

Tabela 2 – Resultados - Parte 1

Autor	Ano	Objetivo	Métodos de detecção	Conjunto de dado	Tempo de coleta	Características	Dados públicos
ESTER et al.	1996	É apresentado o novo algoritmo de agrupamento DBSCAN, confiando em uma noção baseada em densidade de *clusters* que é projetada para descobrir *clusters* de forma arbitrária. O DBSCAN requer apenas um parâmetro de entrada e suporta o usuário na determinação de um valor apropriado para ele.	*Clustering-based*	Híbrido (reais/sintéticos)	N/I	Benchmark SEQUOIA 2000	Não
LEE et al.	1999	A pesquisa centra-se em métodos automáticos para a construção de modelos de detecção de intrusão.	*Classification-based*	Sintético (simulação)	7 semanas	DARPA 1998 - MIT Lincoln Laboratory	Sim
ELEAZAR	2000	É apresentada uma técnica para detectar anomalias sem treinamento em dados normais, através de um método para detectar anomalias usando um conjunto de dados que contém grande número de elementos normais e relativamente poucas anomalias.	*Statistical*	Híbrido (reais/sintéticos)	1) 5 semanas de dados e 3 semanas de teste 2) 15 meses	1) DARPA 1999 - MIT Lincoln Laboratory 2) Stephanie Forrest's - University of New Mexico	Sim [DARPA 1999]
LEE	2001	É proposta a utilização de várias medidas teóricas de informação; entropia condicional, entropia condicional relativa, ganho de informação e custo de informação para a detecção de anomalia. Essas medidas podem ser usadas para descrever as características de um conjunto de dados de auditoria e sugerir os modelos apropriados de detecção de anomalia a serem construídos e explicam a performance do(s) modelo(s).	*Information Theory*	Híbrido (reais/sintéticos)	N/I	1) University of New Mexico sendmail system call data 2) MIT Lincoln Lab (DARPA 1999) sendmail BSM data and MIT Lincoln Lab tcpdump data.	Sim [DARPA 1999]
PORTNOY; ELEAZAR; STOLFO	2001	É apresentado um novo tipo de algoritmo de detecção de intrusão baseado em *cluster*, que treina usando dados não marcados para detectar novas intrusões. Esse método é capaz de detectar tipos diferentes de intrusões, mantendo uma baixa taxa de falso positivo.	*Clustering-based*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim

Tabela 3 – Resultados - Parte 2

Autor	Ano	Objetivo	Métodos de detecção	Conjunto de dado	Tempo de coleta	Características	Dados públicos
NONG; QIANG	2001	É apresentada uma técnica de detecção de intrusão baseada em uma estatística de teste do chi-square.	*Statistical*	Sintético (simulação)	N/I	Dados de auditoria do MIT Lincoln Laboratory	Sim
HAWKINS et al.	2002	Neste artigo foi utilizado um replicador de rede neurais (RNNs) para fornecer uma medida da superação de registros de dados. O desempenho das RNNs é avaliado usando uma medida de pontuação classificada. A eficácia das RNNs para detecção de valores anormais é demonstrada em dois bancos de dados disponíveis publicamente.	*Classification-based*	Sintético (simulação)	1) 5 semanas 2) N/I	1) KDD Cup 1999 2) Wisconsin breast cancer data	Sim
KRÜGEL; TOTH; ENGIN	2002	É apresentado um sistema de detecção de intrusão que usa detecção de anomalia estatística para encontrar ataques *Remote-to-Local*, direcionados à serviços de rede essenciais.	*Statistical*	Sintético (simulação)	7 semanas	DARPA 1998 - MIT Lincoln Laboratory	Sim
GUAN; GHORBANI; JOHN	2003	É proposto um algoritmo de agrupamento baseado no *K-means*, chamado de *Y-means*, para detecção de intrusão. O *Y-means* supera duas falhas que existem no *K-means*: número de *clusters* de dependência e degeneração. Com o *Y-means* um conjunto de dados pode ser particionado automaticamente em um número apropriado de *clusters*.	*Clustering-based*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim
HELLER et al.	2003	É apresentado um Sistema de Detecção de Intrusão Baseado em Host (IDS) que monitora os acessos ao Registro do Microsoft Windows usando o RAD *(Registry Anomaly Detection)*.	*Classification-based*	Real	2 semanas	Dados obtidos à partir do uso de computador com Windows NT 4.0	Não
KRÜGEL	2003	Para mitigar falhas, foi proposto um esquema de classificação de eventos baseado em redes *bayesianas*. Essas redes melhoram a agregação de diferentes saídas do modelo e permitem incorporar informações adicionais.	*Classification-based*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim

Tabela 4 – Resultados - Parte 3

Autor	Ano	Objetivo	Métodos de detecção	Conjunto de dado	Tempo de coleta	Características	Dados públicos
NOBLE et al.	2003	São apresentados dois métodos para a detecção de anomalia baseada em gráficos que foram implementados usando o sistema Subdue. O primeiro, detecção de subestrutura anômala, procura por subestruturas específicas e incomuns dentro de um gráfico. No segundo método, detecção de subgrafos anômalos, o gráfico é particionado em conjuntos distintos de vértices, onde cada um dos quais é testado contra os outros para encontrar padrões incomuns.	*Information Theory*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim
RAMADAS; OSTERMANN; TJADEN	2003	É apresentada uma abordagem baseada em *Map Self-Organizing* para detectar comportamentos de rede anômala desenvolvidos para o INBOUNDS *(Integrated Network Based Ohio University Network Detective Service).*	*Classification-based*	Sintético (simulação)	1) 7 semanas 2) N/I	1) DARPA 1998 - MIT Lincoln Laboratory 2) Construção de um conjunto de dados de conexões HTTP e DNS	Sim [DARPA 1998]
SHYU; CHANG	2003	É proposto um novo esquema que utiliza o classificador robusto de componentes principais (PCA) em problemas de detecção de intrusão, onde os dados de treinamento podem ser não supervisionados.	*Statistical*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim
THOTTAN; JI	2003	Primeiramente são examinados os métodos de detecção de anomalia e descrito em detalhes uma técnica estatística de processamento de sinal baseada na detecção abrupta de mudanças.	*Statistical*	Real	N/I	Rede corporativa da Lucent Technologies e a rede do campus Bell Laboratories	Não
LEUNG; LECKIE	2005	É apresentado um novo algoritmo de agrupamento baseado em densidade e baseado em "grade", que é adequado para a detecção de anomalia não supervisionada. Usando essa técnica, o sistema pode ser treinado com dados não marcados e é capaz de detectar ataques anteriormente "não vistos".	*Clustering-based*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim

Tabela 5 – Resultados - Parte 4

Autor	Ano	Objetivo	Métodos de detecção	Conjunto de dado	Tempo de coleta	Características	Dados públicos
PETROVI; ALVAREZ; CARB	2006	Uma nova estratégia de rotulagem de *clusters* foi proposta para aplicação em um sistema de detecção de intrusão (IDS). Essa estratégia combina a computação do índice do agrupamento e a comparação dos diâmetros dos *clusters*.	*Clustering-based*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim
GERHARD; LI; CARLE	2007	É apresentada uma nova abordagem de *Network Data Mining* que aplica o algoritmo de agrupamento *K-means* para caracterizar conjuntos de dados extraídos de registros de fluxo. Os dados são divididos em *clusters* de intervalos de tempo de tráfego normal e anômalo.	*Clustering-based*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim
POOJITHA; NAVEEN; JAYARAMI	2010	É proposta uma abordagem para detecção de intrusão baseada em rede neurais, treinada pelo algoritmo *Back Propagation*. O resultado do teste mostra que a abordagem proposta funciona bem na detecção de diferentes ataques, com precisão e baixo índice de falso positivo.	*Classification-based*	Sintético (simulação)	1) 7 semanas 2) 5 semanas de dados e 3 semanas de teste 3) 5 semanas	1) DARPA 1998 2) DARPA 1999 3) KDD Cup 1999	Sim
SU	2011	É proposto um método para identificar ataques de inundação em tempo real, com base na detecção de anomalias por classificadores de KNN *(K-nearest-neighbor)* ponderados geneticamente. Além disso, um algoritmo de *clustering* não supervisionado é aplicado para substituir todas as instâncias no conjunto de dados de amostragem com centroides, reduzindo significativamente a despesa de tempo em treinamento e identificação de anomalia em tempo real.	*Clustering-based*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim
PAPALEXAKIS; BEUTEL; STEENKISTE	2012	É analisada a eficácia da utilização de dois algoritmos de *co-clustering* diferentes para ambas as conexões de *cluster*, apontando como detectar conexões anômalas fora de um grande conjunto de dados, sem treinar os algoritmos.	*Clustering-based*	Sintético (simulação)	5 semanas	KDD Cup 1999	Sim

Tabela 6 – Resultados - Parte 5

Autor	Ano	Objetivo	Métodos de detecção	Conjunto de dado	Tempo de coleta	Características	Dados públicos
SYARIF; PRUGEL-BENNETT; WILLS	2012	São descrita as vantagens de usar a abordagem de detecção de anomalia na técnica de detecção de uso indevido, na detecção de intrusões ou ataques de rede desconhecidos. Também investigado o desempenho de vários algoritmos de agrupamento quando aplicado à detecção de anomalia.	*Clustering-based*	Sintético (simulação)	5 semanas	NSL-KDD (uma versão aprimorada do conjunto de dados KDD Cup 1999)	Sim
MOHIUDDIN; MAHMOOD	2013	É fornecida uma nova abordagem não supervisionada para detectar *outliers*, usando um algoritmo de agrupamento *k-means* modificado. Os *outliers* detectados são removidos do conjunto de dados para melhorar a precisão do *cluster*.	*Clustering-based*	Real	N/I	UCI ML Repository	Não
YANG et al.	2013	É proposto um IDS baseado em regras para redes SCADA baseadas no protocolo IEC/104, que inclui abordagens baseadas em assinaturas e baseadas em modelos.	*Classification-based*	Híbrido (reais/sintéticos)	N/I	Usou dados reais para o tráfego de rede e fez modificações sintéticas à partir de tais dados.	Não
AHMED; MAHMOOD	2014	É introduzido o conceito de anomalia coletiva para análise de tráfego de rede. Foi utilizado uma variante do algoritmo *K-means*, o algoritmo *X-means*, para agrupar o tráfego de rede e detectar ataques DoS.	*Clustering-based*	Sintético (simulação)	7 semanas	DARPA 1998 - MIT Lincoln Laboratory	Sim
AMBUSAIDI et al..	2014	É proposto um algoritmo de detecção de intrusão baseado no pressuposto de que a intrusão se comporta de forma diferente do tráfego de rede normal.	*Information Theory*	Sintético (simulação)	5 semanas	NSL-KDD (uma versão aprimorada do conjunto de dados KDD Cup 1999)	Sim

Tabela 7 – Resultados - Parte 6

Autor	Ano	Objetivo	Métodos de detecção	Conjunto de dado	Tempo de coleta	Características	Dados públicos
DELJAC; RANDI; GORDAN	2015	É proposta uma solução complementar para melhorar o desempenho dos sistemas existentes na detecção de falhas. A abordagem baseia-se em um método alternativo baseado em um detector de diagnóstico e estatística híbrido de dois estágios.	*Classification-based*	Real	1 ano	Rede Banda Larga da Croatian T-HT Telecom.	Não
HE et al.	2015	É proposto um sistema de detecção de anomalia em tempo real baseado em *"storm"*, incluindo todos os módulos necessários e todas as teorias usadas neste sistema. O processamento de dados baseia-se em *Hadoop* e *Storm*, usando o *Hadoop* para analisar dados e o *Twitter Storm* para detectar anomalia de tráfego de rede em tempo real.	*Clustering-based*	Real	N/I	2000 portas de acesso de empresas em algumas províncias do sul da China.	Não
PAJOUH; DASTGHAIBYFARD; HASHEMI	2015	Este artigo propõe um novo modelo de classificação de duas camadas com base em abordagens de aprendizado de máquinas *Naive Bayes*.	*Clustering-based*	Sintético (simulação)	5 semanas	NSL-KDD (uma versão aprimorada do conjunto de dados KDD Cup 1999)	Sim
BHUYAN; BHATTACHARYYA; KALITA	2016	É apresentada uma abordagem *multi-step outlier-based* para a detecção de anomalias no tráfego em toda a rede, usando a técnica de seleção de recurso com base em entropia para selecionar um subconjunto relevante de recursos não redundante e também uma técnica de agrupamento em árvore para gerar um conjunto de pontos de referência e uma função de pontuação *outlier* para classificar o tráfego de rede recebido para identificar anomalias.	*Clustering-based*	Híbrido (reais/sintéticos)	1) N/I 2) N/I 3) N/I 4) 5 semanas	1) UCI ML Repository 2) Conjunto de dados de intrusão Real-life TUIDS 3) Conjunto de dados de varredura coordenada Real-life TUIDS 4) KDD Cup 1999 e NSL-KDD	Sim [UCI e KDD]
FERNANDES et al.	2016	São apresentados dois sistemas baseados em perfil para a detecção de anomalia, composta por duas etapas: criação de um modelo que caracterize o comportamento normal do tráfego através de dados históricos e detecção de desvio de comportamento com a ativação de alarmes multinível.	*Statistical*	Real	2 meses	Rede da Universidade Federal de Londrina	Não

Tabela 8 – Resultados - Parte 7

Autor	Ano	Objetivo	Métodos de detecção	Conjunto de dado	Tempo de coleta	Características	Dados públicos
GRILL; PEVNY; REHAK	2016	Neste trabalho é proposto suavizar as saídas dos detectores de anomalia por meio do LAMS *(Local Adaptive Multivariate Smoothing)* que visa reduzir uma grande parte dos falsos positivos na detecção de anomalia.	*Statistical*	Real	N/I	Rede da Universidade Técnica Checa em Praga	Não
MOUSTAFA; SLAY	2016	É discutida a análise e avaliação do conjunto de dado UNSW-NB15. Uma parte deste conjunto de dados é dividida em um conjunto de treinamento e conjunto de testes para examinar este conjunto de dados. Os conjuntos de treinamento e teste são analisados em três aspectos: a fase de análise estatística, a fase de correlação de características e a fase de avaliação de complexidade.	*Statistical*	Sintético (simulação)	31 horas	UNSW-NB15	Sim
GARG; BATRA	2017	É proposta uma técnica de detecção de anomalia robusta, usando agrupamento *Fuzzified Cuckoo based Clustering Technique* (F-CBCT). A técnica opera em duas fases: treinamento e detecção.	*Clustering-based*	Sintético (simulação)	1) N/I 2) 5 semanas	1) UCI ML Repository 2) NSL-KDD (uma versão aprimorada do conjunto de dados KDD Cup 1999)	Sim
KIM et al.	2017	Foi desenvolvida uma técnica de detecção de anomalia *on-line* com duas considerações importantes: disponibilidade de atributos de tráfego durante o tempo de monitoramento e escalabilidade computacional para transmissão de dados.	*Learning-based*	Sintético (simulação)	5 semanas	KDD Cup 1999 e NSL-KDD (uma versão aprimorada do conjunto de dados KDD Cup 1999)	Sim
VIDAL et al.	2017	É introduzida uma estrutura de correlação de alerta que visa complementar as várias deficiências nos sistemas de correlação de alerta e NIDS *(Network Intrusion Detection Systems)* disponíveis na literatura.	*Statistical*	Real	1 ano	Data Center of the University Complutense of Madrid	Não

5 Conclusões e Trabalhos Futuros

O uso de conjuntos de dados sintéticos para validação de técnicas de detecção de anomalias apresenta diversos desafios. O objetivo deste trabalho foi investigar e discutir os resultados obtidos após análise dos artigos contidos no *survey* de Ahmed, Mahmood e Hu (2016).

No decorrer do trabalho, foi discutido brevemente sobre os conceitos básicos relacionados à gerência de rede e introduzido o tema de detecção de anomalias, expondo as causas e os tipos, bem como os diferentes métodos existentes.

Depois de iniciado o tema detecção de anomalias, foram realizadas buscas no Google Scholar por trabalhos relacionados ao conteúdo, buscando por aqueles que foram validados usando algum conjunto de dado, fosse ele real, sintético ou híbrido. Com os trabalhos separados e avaliados, foram então construídas as tabelas com as principais características obtidas durante o estudo de cada um dos artigos.

Alguns números adquiridos através das Tabelas 2, 3, 4, 5, 6, 7 e 8, demonstram que 64% dos trabalhos foram validados com o auxílio de conjuntos de dados sintéticos, 22% utilizando conjuntos de dados reais e 14% com o uso de conjuntos de dados híbridos, o que reforça ainda mais que a maioria dos trabalhos propostos foram validados com o auxílio de conjuntos sintéticos. Segundo Moustafa e Slay (2016), a validação usando os conjuntos de dados de referência existentes como KDD 1999 e o NSL KDD não reflete resultados satisfatórios devido a três problemas principais: (1) a falta de estilos modernos de ataque, (2) a falta de cenários de tráfego normal e (3) uma distribuição diferente de conjuntos de treinamento e teste.

Aprofundando ainda mais, nota-se que os trabalhos publicados a partir de 2015 representam um total de 63% de todos os trabalhos que foram validados com o auxílio de conjuntos de dados reais, indicando uma possível tendência de que os trabalhos mais recentes sejam validados da maneira mais realística possível. Mesmo assim, ainda há trabalhos que foram publicados após 2015, sendo validados utilizando conjuntos de dados sintéticos como o trabalho de Garg e Batra (2017) e Kim et al. (2017), que usaram o conjunto NSL-KDD, uma versão aprimorada do KDD Cup 1999. Nehinbe (2011) reforça que é importante que tais métodos sejam validados da maneira mais realística possível, ou seja, usando conjuntos reais ou conjuntos que se assemelham aos conjuntos de dados reais. Entretanto, essa afirmação é relativa, pois não é possível afirmar que o uso de conjuntos de dados sintéticos não possam refletir resultados satisfatórios.

Como sugestão de trabalhos futuros, pode-se propor formas de padronizar novas propostas de conjuntos de dados usados para a validação de técnicas de detecção de

anomalias, aprofundar a investigação da relação entre técnicas de detecção (*classification-based*, *clustering-based*, *information theory*, *learning-based* e *statistical*) e conjunto de dados usados para a validação e, por fim, comparar a qualidade e o impacto de trabalhos validados usando dados sintéticos e dados reais.

Por fim, espera-se que o presente trabalho seja utilizado para alertar a comunidade acadêmica sobre o uso exacerbado de conjuntos de dados sintéticos obsoletos, como o KDD e DARPA, usados na validação de métodos de detecção de anomalias, pois estes muitas vezes não refletem a realidade, impactando diretamente na eficiência de tais métodos se adotados em ambientes reais. Para isso, algumas soluções possíveis são: buscar conjuntos de dados sintéticos mais atuais, utilizar conjuntos de dados híbridos para mesclar dado real e dado sintético ou então optar por conjuntos de dados reais ou aqueles que se assemelham ao conjunto de dado real.

Referências

ABREU, F. R.; PIRES, H. D. Gerência de Redes. 2004. Citado na página 17.

AHMED, M.; MAHMOOD, A. N. Network Traffic Analysis based on Collective Anomaly Detection. p. 1141–1146, 2014. Citado na página 38.

AHMED, M.; MAHMOOD, A. N.; HU, J. A survey of network anomaly detection techniques. Elsevier, v. 60, p. 19–31, 2016. ISSN 10958592. Citado 5 vezes nas páginas 13, 25, 26, 27 e 41.

AMBUSAIDI, M. A. et al. Intrusion detection method based on nonlinear correlation measure. v. 8, 2014. Citado na página 38.

ANDERSON, J. P. Computer security threat monitoring and surveillance. p. 56, 1980. Citado na página 21.

BALCI, O. Verification Validation and Accreditation of Simulation Models. In: . Washington, DC, USA: [s.n.], 1997. (WSC '97), p. 135–141. ISBN 0-7803-4278-X. Citado na página 28.

BARFORD, P. et al. A Signal Analysis of Network Traffic Anomalies. 2002. Citado 2 vezes nas páginas 13 e 22.

BARTOS, K.; REHAK, M.; KRMICEK, V. Optimizing flow sampling for network anomaly detection. In: . [S.l.: s.n.], 2011. p. 1304–1309. ISSN 2376-6492. Citado na página 13.

BENSO, A. C. Sistema de Detecção de Intrusão baseado em Métodos Estatísticos para Análise de Comportamento. 2003. Citado na página 28.

BHUYAN, M. H.; BHATTACHARYYA, D. K.; KALITA, J. K. A multi-step outlier-based anomaly detection approach to network-wide traffic. Elsevier Inc., 2016. Citado na página 39.

BUENO, E. M. Monitoramento de Redes de Computadores com uso de ferramentas de software livre. 2012. Citado 4 vezes nas páginas 7, 14, 15 e 16.

CELENK, M. et al. Predictive Network Anomaly Detection and Visualization. v. 5, n. 2, p. 288–299, 2010. ISSN 1556-6013. Citado na página 13.

CHAVAN, S. S.; MADANAGOPAL, R. Generic SNMP Proxy Agent Framework for Management of Heterogeneous Network Elements. 2008. Citado 2 vezes nas páginas 7 e 17.

CHRUN, D. Model-based support for information technology security decision making. p. 1–277, 2011. Citado na página 29.

CincoDías. *Telefónica sufre un 'hackeo' masivo y el CNI hace saltar las alarmas: cómo y a quién afecta*. 2017. Disponível em: <http://cincodias.elpais.com/cincodias/2017/05/12/companias/1494585502_908236.html>. Citado na página 12.

CLAFFY, K. C. Internet traffic characterization. 1994. Citado na página 13.

COLE, E. *Network security bible*. [S.l.]: John Wiley & Sons, 2011. v. 768. Citado na página 20.

COMER, D. *Interligação de Redes com TCP/IP: Princípios, Protocolos e Arquitetura*. [S.l.]: Elsevier Brasil, 2015. v. 6. Citado na página 17.

CRISTINA, A.; C, D. A. P. Uma metodologia para caracterização do tráfego de redes de computadores: Uma aplicação em detecção de anomalias. 2011. Citado na página 12.

DELJAC, Z.; RANDI, M.; GORDAN, K. Early detection of network element outages based on customer trouble calls. v. 73, p. 57–73, 2015. Citado 3 vezes nas páginas 29, 31 e 39.

ELEAZAR, E. Anomaly Detection over Noisy data using Learned Probability Distributions. 2000. Citado na página 34.

ESTER, M. et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 1996. Citado na página 34.

ESTEVEZ-TAPIADOR, J. M.; GARCIA-TEODORO, P.; DIAZ-VERDEJO, J. E. Anomaly detection methods in wired networks: a survey and taxonomy. v. 27, n. 16, p. 1569–1584, 10 2004. ISSN 01403664. Citado 3 vezes nas páginas 19, 20 e 21.

FERNANDES, G. et al. Network anomaly detection using IP flows with Principal Component Analysis and Ant Colony Optimization. Elsevier, v. 64, p. 1–11, 2016. ISSN 1084-8045. Citado 3 vezes nas páginas 29, 31 e 39.

G1. *Incêndio em prédio da Algar deixa Uberlândia sem telefone e internet*. 2016. Disponível em: <http://g1.globo.com/minas-gerais/triangulo-mineiro/noticia/2016/05/incendio-em-predio-da-algar-deixa-uberlandia-sem-telefone-e-internet.html>. Citado na página 12.

G1. *Site da Anatel caiu após ser alvo de ataque hacker - notícias em Tecnologia e Games*. 2016. Disponível em: <http://g1.globo.com/tecnologia/noticia/2016/04/site-da-anatel-caiu-apos-ser-alvo-de-ataque-hacker.html>. Citado na página 12.

GARG, S.; BATRA, S. Fuzzified Cuckoo based Clustering Technique for Network Anomaly Detection R. Elsevier Ltd, v. 0, p. 1–20, 2017. ISSN 0045-7906. Citado 2 vezes nas páginas 40 e 41.

GERHARD, M.; LI, S.; CARLE, G. Traffic Anomaly Detection Using K-Means Clustering. 2007. Citado na página 37.

GOGOI, P. et al. A Survey of Outlier Detection Methods in Network Anomaly Identification. n. March, 2011. Citado na página 13.

GRILL, M.; PEVNY, T.; REHAK, M. Reducing false positives of network anomaly detection by local adaptive multivariate smoothing. Elsevier Inc., v. 1, p. 1–15, 2016. ISSN 0022-0000. Citado na página 40.

GUAN, Y.; GHORBANI, A. A.; JOHN, S. Y-means : A Clustering Method for Intrusion Detection. v. 2003, p. 1083–1086. Citado na página 35.

HAINES, J. W. *MIT Lincoln Laboratory: DARPA Intrusion Detection Evaluation*. 2017. Disponível em: <https://www.ll.mit.edu/ideval/data/2000data.html>. Citado na página 30.

HAWKINS, S. et al. Outlier Detection Using Replicator Neural Networks. p. 170–180, 2002. Citado na página 35.

HE, G. et al. A Real-time Network Traffic Anomaly Detection System based on Storm. p. 153–156, 2015. Citado na página 39.

HE, L.; YU, S.; LI, M. Anomaly Detection Based on Available Bandwidth Estimation. In: . [S.l.: s.n.], 2008. p. 176–183. Citado na página 13.

HELLER, K. A. et al. One Class Support Vector Machines for Detecting Anomalous Windows Registry Accesses. Citado na página 35.

JAIN, A.; MURTY, M.; FLYNN, P. Data Clustering : A Review. v. 31, n. 3, 1999. Citado na página 27.

KIM, J. et al. A Lightweight Network Anomaly Detection Technique. p. 1–5, 2017. Citado 2 vezes nas páginas 40 e 41.

KRÜGEL, C. Bayesian Event Classification for Intrusion Detection. n. Acsac, 2003. Citado na página 35.

KRÜGEL, C.; TOTH, T.; ENGIN, K. Service Specific Anomaly Detection for Network Intrusion Detection. p. 201–208, 2002. Citado na página 35.

LAKHINA, A.; CROVELLA, M.; DIOT, C. Characterization of network-wide anomalies in traffic flows. v. 6, p. 201, 2004. ISSN 00283940. Citado na página 22.

LEE, W. Information-Theoretic Measures for Anomaly Detection. p. 130–143, 2001. Citado 3 vezes nas páginas 28, 30 e 34.

LEE, W. et al. A Data Mining Framework for Building Intrusion Detection Models. 1999. Citado na página 34.

LEUNG, K.; LECKIE, C. Unsupervised Anomaly Detection in Network Intrusion Detection Using Clusters. v. 38, n. January, p. 333–342, 2005. Citado na página 36.

LIM, S. Y.; JONES, A. Network Anomaly Detection System: The State of Art of Network Behaviour Analysis. In: . [S.l.: s.n.], 2008. p. 459–465. Citado na página 21.

MOHIUDDIN, A.; MAHMOOD, A. N. A Novel Approach for Outlier Detection and Clustering Improvement. p. 577–582, 2013. Citado na página 38.

MOUSTAFA, N.; SLAY, J. The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. v. 3555, n. January, p. 0–14, 2016. Citado 3 vezes nas páginas 31, 40 e 41.

NEHINBE, J. O. A critical evaluation of datasets for investigating IDSs and IPSs researches. p. 92–97, 2011. Citado 2 vezes nas páginas 13 e 41.

NOBLE, C. C. et al. Graph-Based Anomaly Detection. v. 1, p. 631–636, 2003. Citado na página 36.

NONG, Y.; QIANG, C. An Anomaly Detection Technique Based on a Chi-Square Statistic for Detecting Intrusions Into Information Systems. p. 105–112, 2001. Citado na página 35.

PAJOUH, H. H.; DASTGHAIBYFARD, G.; HASHEMI, S. Two-tier network anomaly detection model : a machine learning approach. n. 2, 2015. ISSN 0925-9902. Citado na página 39.

PAPALEXAKIS, E. E.; BEUTEL, A.; STEENKISTE, P. Network Anomaly Detection using Co-clustering. 2012. Citado 2 vezes nas páginas 23 e 37.

PATCHA, A.; PARK, J.-M. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Elsevier, v. 51, n. 12, p. 3448–3470, 2007. Citado 2 vezes nas páginas 20 e 28.

PERLIN, T.; NUNES, R.; KOZAKEVICIUS, A. D. J. Detecção de Anomalias em Redes de Computadores através de Transformadas Wavelet. v. 35, n. 1, p. 109, 10 2011. ISSN 2376-6492. Citado na página 12.

PETROVI, S.; ALVAREZ, G.; CARB, J. Labelling Clusters in an Intrusion Detection System Using a. v. 00, n. C, p. 1–8, 2006. Citado na página 37.

POOJITHA, G.; NAVEEN, K.; JAYARAMI, P. Intrusion Detection using Artificial Neural Network. 2010. Citado 2 vezes nas páginas 29 e 37.

PORTNOY, L.; ELEAZAR, E.; STOLFO, S. J. Intrusion Detection with Unabeled Data Using Clustering. 2001. Citado na página 34.

PROENCA, M. et al. Baseline to help with network management. p. 158–166, 2006. Citado na página 20.

PROVDANOV, C. C.; FREITAS, E. C. D. *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico*. [S.l.: s.n.], 2013. 276 p. ISSN 1098-6596. ISBN 9788577171583. Citado na página 25.

RAMADAS, M.; OSTERMANN, S.; TJADEN, B. Detecting Anomalous Network Traffic with Self-organizing Maps. p. 36–54, 2003. Citado na página 36.

RINGBERG, H. et al. Sensitivity of PCA for traffic anomaly detection. v. 35, n. 1, p. 109, 2007. ISSN 01635999. Citado na página 22.

ROUGHAN, M. et al. IP forwarding anomalies and improving their detection using multiple data sources. p. 307, 2004. Citado na página 22.

SAITO, J. T.; MADEIRA, E. Um Modelo de Gerenciamento de Redes de Telecomunicações Utilizando a Plataforma CORBA. 2001. Citado na página 15.

SHARAFALDIN, I. et al. Towards a Reliable Intrusion Detection Benchmark Dataset. v. 2017, n. 1, p. 177–200, 2017. ISSN 2445-9739. Citado na página 29.

SHIRAVI, A. et al. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Elsevier Ltd, v. 31, n. 3, p. 357–374, 2012. ISSN 01674048. Citado 2 vezes nas páginas 29 e 30.

SHON, T.; MOON, J. A hybrid machine learning approach to network anomaly detection. v. 177, n. 18, p. 3799–3821, 2007. ISSN 00200255. Citado na página 23.

SHYU, M.-l.; CHANG, L. A Novel Anomaly Detection Scheme Based on Principal Component Classifier. 2003. Citado na página 36.

STALLINGS, W. *Redes e sistemas de comunicação de dados*. [S.l.]: Elsevier, 2005. Citado na página 14.

SU, M.-y. Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification. Elsevier, v. 34, n. 2, p. 722–730, 2011. ISSN 1084-8045. Citado na página 37.

SYARIF, I.; PRUGEL-BENNETT, A.; WILLS, G. Unsupervised clustering approach for network anomaly detection. 2012. Citado na página 38.

TAN, Z. et al. Enhancing big data security with collaborative intrusion detection. v. 1, n. 3, p. 27–33, 2014. ISSN 23256095. Citado na página 26.

THOTTAN, M.; JI, C. Anomaly Detection in IP Networks. 2003. Citado 6 vezes nas páginas 12, 19, 20, 21, 31 e 36.

VIDAL, J. M. et al. Alert Correlation Framework for Malware Detection by Anomaly-based Packet Payload Analysis. Elsevier Ltd, 2017. ISSN 1084-8045. Citado na página 40.

WANG, H. et al. Detection Network Anomalies Based on Packet and Flow Analysis. In: . [S.l.: s.n.], 2008. p. 497–502. Citado na página 13.

YANG, Y. et al. Rule-Based Intrusion Detection System for Scada Networks. p. 2–5, 2013. Citado na página 38.

YE, N. et al. Multivariate Statistical Analysis of Audit Trails for Host-Based Intrusion Detection. v. 51, p. 810–820, 2002. Citado na página 28.

ZARPELÃO, B. B. Detecção de Anomalias em Redes de Computadores. 2010. Citado 5 vezes nas páginas 12, 13, 18, 19 e 23.

ZHANG, W.; YANG, Q.; GENG, Y. A Survey of Anomaly Detection Methods in Networks. p. 9–11, 2009. Citado na página 19.

ZHANI, M. F.; ELBIAZE, H.; KAMOUN, F. Analysis of prediction performance of training-based models using real network traffic. In: . [S.l.: s.n.], 2008. p. 472–479. Citado na página 13.